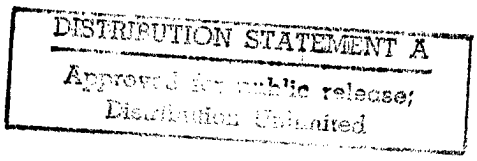
Report IBM #DABT63-94-C0042 Final

Robust Models and Features for Speech Recognition

David Nahamoo, Principal Investigator
IBM Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

March 13, 1998
Final Report covering 9/24/93 to 12/31/97



Robust Models and Features for Speech Recognition

David Nahamoo, Principal Investigator
Contract Period: September 24 1993 to December 31 1997

March 13, 1998

1 Technical Objectives

The goal of this project were: (1) to develop algorithms for achieving significant improvements in accuracy over current state-of-the-art speech recognition systems, (2) to implement these algorithms and assemble them into a prototype for a hands-on demonstration and (3) to test these algorithms for the transcription of "real-world" speech like radio/TV broadcast shows.

2 Methodology

The primary aim of the project was to develop algorithms for addressing robustness problems in large vocabulary, speaker-independent, continuous speech recognition. Problems given particular focus are (i) robustness with respect to channel (noise and microphone) change and variation, (ii) effective modeling of conversational speech variability, (iii) enhanced modeling of pronunciation variability in speaker-independent speech recognition, and (iv) optimal utilization of the training data available for modeling.

We attempted to tackle the robustness of speech recognition using two-pronged approach. First, speaker independent features and models are developed that provide a certain degree of robustness to variation across different speakers and, to a lesser extent, variations in noise characteristics. This is exemplified by our linear discriminant based features, Gaussian mixture based hidden Markov models, and rank-based decoding strategy. By relying only on the rank order of probabilities produced by Gaussian mixture models rather than on their absolute values, we can increase the total number of model parameters and thereby improve performance without sacrificing robustness. In particular, we employ such ideas towards improved modeling of the HMM output probability distributions. We also currently use context-dependent modeling built using decision networks.

Further, we addressed robustness and adpatability using data-driven methods. We addressed the adaptability of speech recognition systems to various environments and speakers by developing

methods for unsupervised adpatation that use a very small amount of data from the new environment to quickly adapt the system parameters to more closely match new environments and speakers. Since these methods are unsupervised - that is, they do not require the knowledge of the correct sentences that were spoken, relying on the results of an intial decoding pass instead - they can be applied in a wide variety of cirumstances.

We improved the robustness of the recognition system based on model combination techniques. These techniques use a very small segment (about one second) of the background signal to modify the model parameters so as to reflect the new signal condition more accurately. Thus it is consistent with the above adaptation approach and provides the possibility of natural integration of the two approaches. With this technique we expect to filter out a good deal of degradation caused by microphone variations and environment changes.

The accuracy of a recognizer is heavily tied to its robustness to microphone and background noise variations, as well as other dynamic characteristics commonly occurring in spontaneous conversational speech. We improve the robustness of the recognition system based on training systems using data with a variety of background noise and channel conditions. This gives a model that is very robust but not especially matched to any particular environment or background noise. Unsupervised adaptation (to speaker, environment etc.) is then used to improve the accuracy of the system. Several new unsupervised adaptation schemes like Clustered Transformations [7, 12, 13], Adaptation by Correlation [10, 11], and Speaker Dependent Variances have been developed for this purpose.

Another part of the project dealt with the class of problems that can be tackled through enhanced speech recognition modeling. Specifically, we employ automatic context dependent model techniques based on the concept of *Decision Trees/Networks* and These tools lead to powerful models for handling speaker and task variations. These models make very efficient use of available data and substantially outperform simpler models that existed in state-of-the-art recognition systems at the start of this project.

All of the above algorithms were tested on the Hub and relevant Spokes of the Speaker Independent Wall Street Journal database in 1994, the Marketplace database in 1995, and the Broadcast news database in 1996. A real-time implementation of our recognizer for read NAB news was demonstrated at the ARPA workshop in 1995.

3 Accomplishments

3.1 Accomplishments in 1994-1995 Funding Period

During this funding period the main focus was of our effort was on building and improving our baseline speech recognition system for recognizing "read" business news articles. Secondarily, we

also tried to improve the recognition accuracy in the presence of background noise. These activities were geared towards the 1994 DARPA Hub1 NAB news task and the 1994 DARPA Spoke10 noise task. In the Hub1 evaluation in November 1994, IBM's system ranked second among all the participants, a significant achievement considering the fact that it was the first time that IBM had participated in a DARPA speech recognition evaluation. Furthermore, in the Spoke10 noise task IBM's recognizer was placed favorably above all the other participants, validating the efficiency and effectiveness of our noise compensation algorithms.

In addition, as part of this effort, we integrated the techniques developed here with our other independent work on decoder design and produce a prototype for demonstrating the feasibility of the developed techniques. This demo system benefits from our experience in developing real-time client-server and standalone systems that are part of the IBM speech recognition product line. To accomplish this, significant time and effort was spent in reducing the computational, storage and memory requirements of our recognizer. As a result of this work, a real-time prototype system for NAB news transcription was demonstrated at the ARPA workshop in Austin in January 1995.

All these accomplishments were made possible through a sequence of activities: 1. A base acoustic model was trained using all the utterances from the WSJ0+1 corpus distributed by LDC. System tuning and algorithmic improvements on this base system resulted in a system giving 30% relative gains in accuracy on the 1993 Hub0 WSJ evaluation test.

2. Besides the official 20000 word vocabulary, we also built a 64000 word vocabulary. Language models for this vocabulary were built from a combination of Wall Street Journal data available from the LDC and a private collection.

3. We conducted several experiments to establish the relative importance of various training strategies.

4. Context-dependent models using decision trees and networks were developed using the NAB News task as a testbed [1]

5. A double-rotation based Linear discriminant analysis (LDA) scheme to generate feature vectors was shown to provide superior accuracy compared to using first and second order differences of cepstral features [3, 4].

6. Context dependent distribution modeling was achieved using ranks. Methods for using the rank order for determining the output probability values in a hidden Markov model were explored [3, 4].

7. The recognizer was modified to to produce the top N best sentence hypotheses instead of the just the best. This was used to as the input to a rescoring program which use continuous parameter hidden Markov models. This allowed a comparison of rank-based decoding vs. continuous parameter decoding and the former was marginally better.

8. We implemented a variant of Paralled Model Combination for noise compensation. This

was shown to give significant improvement in the presence of car noise [5].

9. The components that were developed were integrated with IBM's pre-existing speech decoder and a real-time demonstration on large vocabulary continuous speech recognition in a speaker independent fashion was presented at the DARPA Spoken Language Systems Technology Workshop in Austin Texas in February 1995.

3.2 Accomplishments in 1995-1996 Funding Period

In June 1995 research emphasis shifted to the transcription of radio and TV broadcast news. In this calendar year we participated in the 1995 dry run for the Hub4 broadcast news transcription task and performed significantly better than the other participants. The algorithmic improvements that made this possible are:

1. A adaptive-music cancellation scheme was designed and implemented to remove music/noise from music/noise corrupted speech. While this lead to perceptually better speech (i.e., was an enhancement scheme for speech), it did not translate to better recognition accuracy [8].

2. Length-constrained HMM models were devised to segment the acoustic data into telephone bandwidth speech, clean speech, pure music and music corrupted speech. This facilitated the use of separate acoustic models for recognizing speech with various background conditions [6].

3. A speaker identification scheme was developed using gaussian mixture models for all the main speakers in Marketplace show. This was used further for speaker adaptation [7].

4. Acoustic models were build for Marketplace shows using Bayesian adaptation of the NAB news task models from the 1994 evaluation [7].

5. MLLR speaker adaptation scheme was implemented.

6. A new speaker adaptation scheme known as Clustered Transformations was developed. The basic idea is to find the "closest N" speakers to a test speaker and build a model based on the training data for these N speakers. This was shown to give significant improvements over standard MLLR adaptation [7, 12, 13].

7. Speaker specific models were used in the ARPA evaluation for known speakers as recognized by our speaker identification algorithms.

8. A sequence of binary classification schemes was used for the segmentation with length-constrained HMMs. This allowed for the use of feature spaces that are particular suited for a particular classification. For example, 24-dimensional cepstral space was used for tele/non-tele segmentation while 60-dimensional LDA space was used for music/non-music segmentation.

3.3 Accomplishments in 1996-1997 Funding Period

Through 1996 and to the present time the focus of our research effort has continued to be the transcription of broadcast news. The accomplishments during this period are as summarized below:

1. Processed the first 50 hours of BN training data distributed by LDC to develop the acoustic models used in the 1996 Hub4 evaluation.

2. We participated in both the UE and PE evaluations.

3. The code-base for deleted interpolation 4-gram language models was developed. While 4-gram language models by themselves did not give any gain, in conjunction with trigram language models (used in a mixture) gave small gains in recognition performance.

4. A new speaker adaptation scheme, Adaptation By Correlation (ABC) was developed. Unlike MLLR, ABC is fairly robust when there is very little adaptation data. It is based on the following idea: for those gaussians that have sufficient adaptation data the adapted gaussians are estimated in the standard fashion; for those that do not, the correlations between the gaussians (obtained from training data) is used to estimate mean shifts [10, 11].

5. We proposed the use of non-parametric densities for the HMM states (in lieu gaussian mixture models). In general, non-parametric densities require large amounts of data to get reliable estimates. However, wavelet-smoothed histograms can be used as density estimates to better model HMM states in some cases. This proposal requires enormous algorithmic improvements before it can make it to practical speech recognizers and is still being investigated currently.

6. We implemented VTL normalization into our front-end processing. Gains in the baseline performance due to VTL vanished after unsupervised speaker adaptation using MLLR. Because of this VTL processing is currently not being used in our BN speech recognizer.

7. We implemented and experimented with Speaker Adaptive Training (SAT) on the WSJ database. Because of the enormous computational and storage requirements of SAT training, this algorithm was not used in the 1996 Hub4 evaluation system.

8. In early 1997 all the 100 hours of BN training data arrived from LDC. Since the transcripts had been changed and corrected (for the already processed shows), the entire training data was processed from scratch. Interestingly, there was very little gain in performance because of the additional 50 hours of training data, implying that additional gains have to come from algorithmic improvements.

9. Experimented with a technique known as speaker dependent variances. The idea is to use the average speaker-dependent variance from the training data on test data *after* the means have been adapted using MLLR or any other technique. This sharpens the models and increases the likelihood leading to better accuracy.

10. Implemented PLP (perceptual linear prediction) based signal processing for feature extraction. This did not seem to improve or degrade performance in clean or noisy data. Marginal robustness to microphone variations was noticed.

11. We conducted a preliminary study on the use of non-parametric models (using wavelet shrinkage - which allows one to store non-parametric models efficiently) that can potentially give

better models and lead to faster decoding (since likelihoods can be obtained by table lookup). The reason for considering this is that the distributions for several of the features a posteriori is non-gaussian. Feature dimension that are well-modeled using classical parametric methods can be left alone leading to hybrid (non-parametric plus parametric) models. In this line of enquiry we have just scratched the surface.

4 Significance

During the course of this contract IBM for the first time started participating in the ARPA evaluations. In the context of these evaluations progress was made from transcribing clean read speech (Wall Street Journal task in 1994) to real world speech (transcription of radio and TV broadcast news in 1997). We believe significant progress has been made in acoustic feature extraction (LDA etc.), acoustic modeling (decision trees, rank probabilities) as well as search (envelope stack search extended to generate N-best lists). All this, along with appropriate training data (supplied by LDC as part of the ARPA program) has enabled us to obtain respectable speech recognition performance on real world speech - 18% word error rate on the 1997 Hub4 evaluation with broadcast news. In comparison on relatively easier data (Marketplace news transcription) our best number in 1995 was 27%. Moreover, in early 1995 performance on the same data (Marketplace) was at 40% word error rate. The algorithms developed in this project for robust features, robust models and rapid unsupervised adaptation have played a significant role in this error-rate reduction. Speech recognizer performance on real data is at that stage where practically useful real-world applications can be built on it e.g., audio indexing, dictation applications, voice navigation etc. However, to make speech the natural human interface to computers, still further algorithmic work needs to be done on improving the accuracy, robustness, and speech.

5 Conclusions

The state-of-art in speech recognition has evolved considerably during the course of this project as evidenced by the nature and complexity of the tasks that are being currently addressed by researchers worldwide. Focus has shifted from rom small-vocabulary "read-speech" tasks like Resource Management to "found speech" tasks like the transcription of broadcast news. During the course of this project we made several significant contributions that made this transition possible for the speech recognition community in general and the IBM system in particular. The key contributions were robust features and models and rapid model and feature adaptation to unknown speakers and environments. Besides, we introduced several techniques that improve the baseline accuracy and allow us to move to ever more complex recognition tasks: LDA double-rotation features, rank-based decoding, decision trees for context clustering, noise compensation, envelope

search, automatic segmentation and clustering of speech (along speakers, channels etc), Adaptation by Correlation, Clustered (speaker) Transformations, Speaker-Dependent Variances etc. Besides developing new techniques we have also taken algorithms developed by other groups around the world and tested their validity and usefulness on other data sets and tasks e.g., Vocal Tract Length normalization, Speaker Adapted Training, Perceptual Linear Prediction features etc. As part of this project a real-time speech recognition system for read financial news was demonstrated at the ARPA workshop in 1995.

6 List of Contract Publications

References

[1] L. R. Bahl et al., "Robust Methods for using Context-Dependent features and models in a continuous speech recognizer", Proc. ICASSP, 1994.

[2] Jerome R. Bellegarda, Peter V. de Souza, David Nahamoo, Mukund Padmanabhan, Michael A. Picheny, Lalit R. Bahl, "Experiments Using Data Augmentation for Speaker Adaptation," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Detroit, Michigan, May 1995, pp.692-695.

[3] L. R. Bahl, S. Balakrishnan-Aiyer, J.R. Bellegarda, M. Franz, P.S. Gopalakrishnan, D. Nahamoo, M. Novak, M. Padmanabhan, M.A. Picheny, S. Roukos, "Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Detroit, Michigan, May 1995, pp.41-44.

[4] L. Bahl, S. Balakrishnan-Aiyer, M. Franz, P.S. Gopalakrishnan, R. Gopinath, M. Novak, M. Padmanabhan, S. Roukos, "The IBM Large Vocabulary Continuous Speech Recognition System for the ARPA NAB News Task," Proceedings of the Spoken Language Systems Technology Workshop, Austin, Texas, January 22-25, 1995, pp.121-126

[5] R.A. Gopinath, M. Gales, P.S. Gopalakrishnan, S. Balakrishnan-Aiyer, M. Picheny, "Robust Speech Recognition in Noise - Performance of the IBM Continuous Speech Recognizer on the ARPA Noise Spoke Task," Proceedings of the Spoken Language Systems Technology Workshop, Austin, Texas, January 22-25, 1995, pp.127-130

[6] Gopalakrishnan, P.S., Gopinath, R., Maes, S., Padmanabhan, M., Polymenakos L., "Transcription of radio broadcast news with the IBM large vocabulary speech recognition system," Proceedings of the DARPA Speech Recognition Workshop, Arden House, Feb 1996.

[7] P. S. Gopalakrishnan, et al., "Acoustic Models Used in the IBM System for the ARPA Hub 4 Task," Proceedings of the DARPA Speech Recognition Workshop, Arden House, Feb 1996.

[8] P. S. Gopalakrishnan, D. Nahamoo, M. Padmanabhan and L. Polymenakos. "Suppressing background music from music-corrupted data in the ARPA Hub4 task", Proceedings of the DARPA Speech Recognition Workshop, Arden House, Feb 1996.

[9] Bakis, R., Chen, S., Gopalakrishnan, P.S., Gopinath, R.., Maes, S., Polymenakos, L., "Transcription of broadcast news shows with the IBM large vocabulary recognition system," Proceedings of the DARPA Speech Recognition Workshop, Chantilly, VA, Feb 1997.

[10] S. Chen and P. DeSouza "Adaptation By Correlation", Proceedings of the DARPA Speech Recognition Workshop, Chantilly, VA, Feb 1997, pp 123-126, 1997.

[11] S. Chen and P. DeSouza, Adaptation by Correlation, Proceedings of Eurospeech 1997.

[12] M. Padmanabhan, L. R. Bahl, D. Nahamoo, M. Picheny, "Speaker Clustering and Transformation for Speaker Adaptation in Speech Recognition Systems", Proceedings of ICASSP 1996, vol II, pp. 701-704, May 1996.

[13] M. Padmanabhan, L. R. Bahl, D. Nahamoo, M. Picheny, "Speaker Clustering and Transformation for Speaker Adaptation in Speech Recognition Systems", IEEE Trans. Speech and Audio Proc. vol. 6, pp. 71-77, Jan 1998.